

The application of annotation models for the construction of databases and tools

Overview and analysis of MPI work since 1994

H. Brugman, P. Wittenburg
MPI Nijmegen



Max Planck Institute for Psycholinguistics



Overview

- 4 generations of models for linguistic annotations since 1994
 - MediaTagger's implicit model (1994)
 - Relational model (1996)
 - Abstract Corpus Model, version 1 (1997/98)
 - ACM, version 2 (2001)
- A critical comparison
- Comparison to ATLAS/AIF



Max Planck Institute for Psycholinguistics



MediaTagger's model

- Based on QuickTime's implicit model
- Unrestricted number of user-definable tiers
- Time dependencies between tiers
- Primitive annotation/tier types, including closed vocabularies



Max Planck Institute for Psycholinguistics



Relational database model

- Entity-Relationship diagram using a formal method
- Carefully normalized set of database tables
- Multiple users share one set of tables
- Reusable, sharable tier types
- Client-server operation with MT using ODBC



Max Planck Institute for Psycholinguistics



Abstract Corpus Model, v1

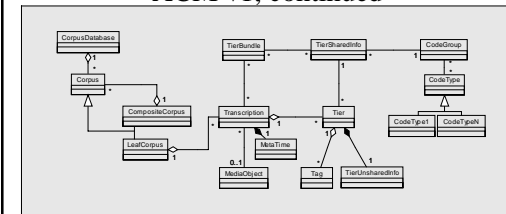
- Main aims:
 - Annotation file format independent
 - Extensible for both tools and file formats
 - 3 layer, OS independent, distributed architecture
 - Internet capable operation, including streaming media
- Object oriented model
 - Set of cooperating *interface* definitions
 - Implemented as remote abstract classes
 - Specific subclasses for each annotation file format



Max Planck Institute for Psycholinguistics



ACM v1, continued



Max Planck Institute for Psycholinguistics



Abstract Corpus Model, v2

- Main problems addressed
 - IMDI metadata integration
 - Support for more complex patterns of annotations
 - Pointing at, instantiation and initialization of ACM objects for resources at varying locations
 - Far better support for tier types



Max Planck Institute for Psycholinguistics



Comparison

| | mt | rdb | acm1 | acm2 |
|--------------------------------|----|-----|------|------|
| Unrestricted # of tiers | v | v | v | v |
| Document metadata | - | + | + | ++ |
| Tier metadata | - | + | + | ++ |
| Concurrent editing of document | | v | v | v |
| File format independence | | | v | v |
| Distributed operation | | | v | v |



Max Planck Institute for Psycholinguistics



Comparison, continued

| | mt | rdb | acm1 | acm2 |
|------------------------------------|----|-----|------|------|
| Partial time alignment | | | v | v |
| Time constrained in parent segment | v | v | | v |
| Times duplicated between tiers | v | v | | |
| Strictly consecutive segments | | | | v |
| Annotations can share times | | | | v |



Max Planck Institute for Psycholinguistics



Comparison, continued

| | mt | rdb | acm1 | acm2 |
|---------------------------------|----|-----|------|------|
| 1-1 symbolic references | | | v | v |
| 1- <i>n</i> symbolic references | | | | v |
| <i>n</i> -1 symbolic references | | | | v |
| Symbolic trees (e.g. syntax) | | | | v |
| Ordered reference annotations | | | | v |
| Structured annotation content | | | v | |



Max Planck Institute for Psycholinguistics



Comparison, continued

| | mt | rdb | acm1 | acm2 |
|---|----|-----|------|------|
| Closed vocabularies | v | v | | v |
| Unicode support | | | v | v |
| Reusable tier types/templates | | v | v | v |
| Reusable document templates | | | | v |
| Location of media resources | - | - | + | ++ |
| Location and instantiation of annotation resource objects | - | + | + | ++ |



Max Planck Institute for Psycholinguistics



ACM and ATLAS/AIF

- Tier concept is lacking
- Explicit time ordering got lost
- References from time aligned annotations could cause conflicts
- Structured annotation content
- AIF and Eudico Annotation Format (EAF)



Max Planck Institute for Psycholinguistics



Conclusions

- Generic model is *contradictio in terminis*
- Explicit modeling is a solid basis for building good tools
 - Integrated toolset for metadata and annotations on basis of IMDI and EUDICO technology (freely available for academic use).
 - Demonstrations on Wednesday and Thursday



Max Planck Institute for Psycholinguistics

